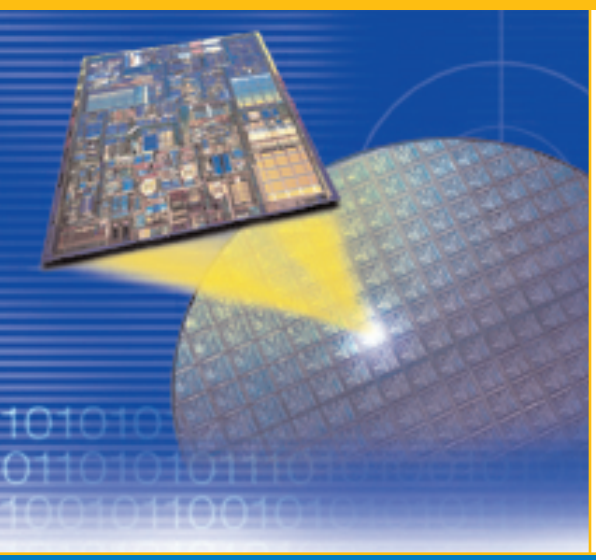
Volume 06 Issue 02 Published, May 16, 2002 ISSN 1535766X



Intel® Technology Journal

Semiconductor Technology and Manufacturing

ETOX™ Flash Memory Technology: Scaling and Integration Challenges

A compiled version of all papers from this issue of the Intel Technology Journal can be found at:
http://developer.intel.com/technology/itj/index.htm

ETOX™ Flash Memory Technology: Scaling and Integration Challenges

Al Fazio, California Technology and Manufacturing, Intel Corp.
Stephen Keeney, California Technology and Manufacturing, Intel Corp.
Stefan Lai, California Technology and Manufacturing, Intel Corp.

Index words: Flash memory, ETOX™, Intel StrataFlash® memory, Moore's Law

ABSTRACT

The 0.13 m flash memory technology that started high-volume manufacturing in the first quarter of 2002 is the eighth generation of flash technology since its first conception and development in 1983. The scaling has been accomplished by improved lithography capability as well as many process architecture innovations. In this paper, the key scaling challenges as well as the key innovations are presented. It is projected that the current planar cell structure can be scaled to the 65nm node. More revolutionary innovations, such as 3D structures, may be required for the 45nm node and beyond. To lower cost further, Intel StrataFlash memory technology has been developed, which stores two bits of information in a single physical memory cell. The scaling innovations also allow for the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology. These integration challenges are also discussed.

INTRODUCTION

The in-system update and non-volatile capabilities of flash memories have enabled it to become the memory of choice for many emerging markets over time, originally as point of sales system configurations, then as PC BIOS components, and today for cell phones and handheld computing devices [1]. Similar to other memory technologies, ETOX™ flash memory scaling follows Moore's law. Figure 1 shows SEM cross-sections of the memory cells for eight generations of flash memory technologies. The memory cell size for the first generation based on 1.5 m lithography was 36 m^2, whereas the cell size for the latest 0.13 m lithography is 0.154 m^2. This represents an over 230 times cell size reduction over the eight generations. In the same period, the memory density for peak volume has increased one thousand fold from 64Kb to 64Mb.

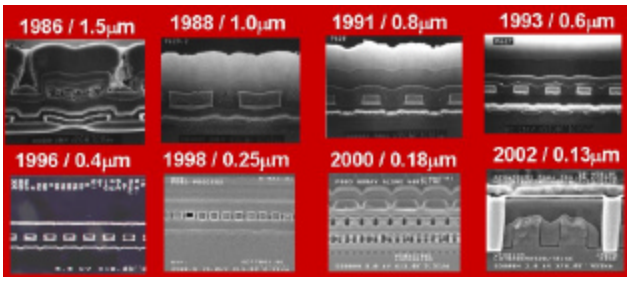


Figure 1: Eight generations of flash technology

Although scaling the flash cell is important to achieve die size reduction or larger memories, the periphery transistors must also be scaled. Scaling the periphery transistors can be achieved by reducing the maximum voltages that need to be supported along with junction engineering and more advanced lithography and etch capabilities. The process architecture innovations and scaling of periphery transistors enables the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology. In this paper we review the key process architecture innovations for scaling, the Intel StrataFlash memory technology and the key innovations required for "wireless Internet on a chip" technology. Table 1 outlines the key innovations for each generation of flash memory.

Table 1: Innovations by technology generation

Technology Node	Key Innovation
1.5 m	Established Flash
1.0 m	Isolation rounding reduction for improved cell gate alignment Cycling reliability established
0.8 m	Recessed LOCOS
0.6 m	Self Aligned Source Scaled Array Field Oxide
0.4 m	Negative Gate Erase Intel StrataFlash memory
0.25 m	Trench Isolation Salicide
0.18 m [2]	Self Aligned floating gate Unlanded Contacts Multiple Periphery Gate Oxides
0.13 m [3]	Channel Erase Dual Trench Dual gate Spacer Wireless Internet on a Chip

FLASH CELL SCALING

Cell size scaling is achieved by scaling critical area components. Each of the key scaling components is described. Figure 2 illustrates cell layout and scaling constraints. A key enabler to scaling is improved line width and space definition through new lithography at each generation. Architecture innovations, such as a number of self-aligned techniques, provide the bulk of the remaining area reduction.

CELL WIDTH (WORDLINE DIRECTION)

The cell width is determined by the simultaneous constraints of isolation pitch (isolation and cell active diffusion); floating gate pitch (endcap, space, and alignment); and contacted metal pitch (contact size, contact and metal space, and alignment). Each of these needs to be scaled in order to scale the cell width.

Isolation Pitch

Two key approaches have been adopted over the last several generations that have enabled continuous pitch scaling. The first is the adoption of a dual isolation scheme where the flash array isolation is decoupled from the periphery isolation so each can be optimized independently. This was first introduced in a local oxidation of a silicon isolation scheme, LOCOS, in the 0.6 m generation. The second key enabler was the introduction of trench isolation at the 0.25 m node, which helped to reduce the active width loss in the device. For the 0.13 m generation, a dual isolation scheme was adopted, now called dual trench, where the array trench was made shallower than the periphery trench for independent optimization. As before with the dual LOCOS scheme, the flash cell can be scaled more aggressively while still meeting the periphery isolation requirements. At each technology node, improved lithography capability is utilized. Additionally, improved gap fill capability of High-Density Plasma (HDP) oxides has been utilized since the 0.18 m technology node.

Floating Gate Pitch

The correct alignment of the floating gate to the active area is a very important cell size determinant, and it becomes more of a constraint as the isolation pitch is scaled and the floating gate isolation is constrained by the lithography minimum space capability. The 0.18 m technology node introduced a new self-aligned scheme (Figure 3, left half) where the floating gate is self-aligned to the isolation using a chemical mechanical polish process. This has been carried forward to the 0.13 m node as well. This self-aligned scheme removes the registration component of the scaling and also allows a sub-lithographic poly space.

Contacted Metal Pitch

Each generation takes advantage of the advances in lithography to scale the contact size and metal pitch. However, the contact alignment to the active area became the constraint at the 0.18 m node, and an UnLanded Contact (ULC) scheme was introduced (Figure 3, right half). In this case, a nitride etch stop layer is deposited below the inter-layer-dielectric oxide to prevent the contact etch punching through the isolation and causing a short to the substrate. This allows the contact to land partially in the isolation and reduces the registration constraint. This ULC scheme is continued in the 0.13 m technology.

CELL HEIGHT (BITLINE DIRECTION)

The cell height is determined by constraints of contact size and contact-to-gate alignment, gate length and drain and source space (source rail width).

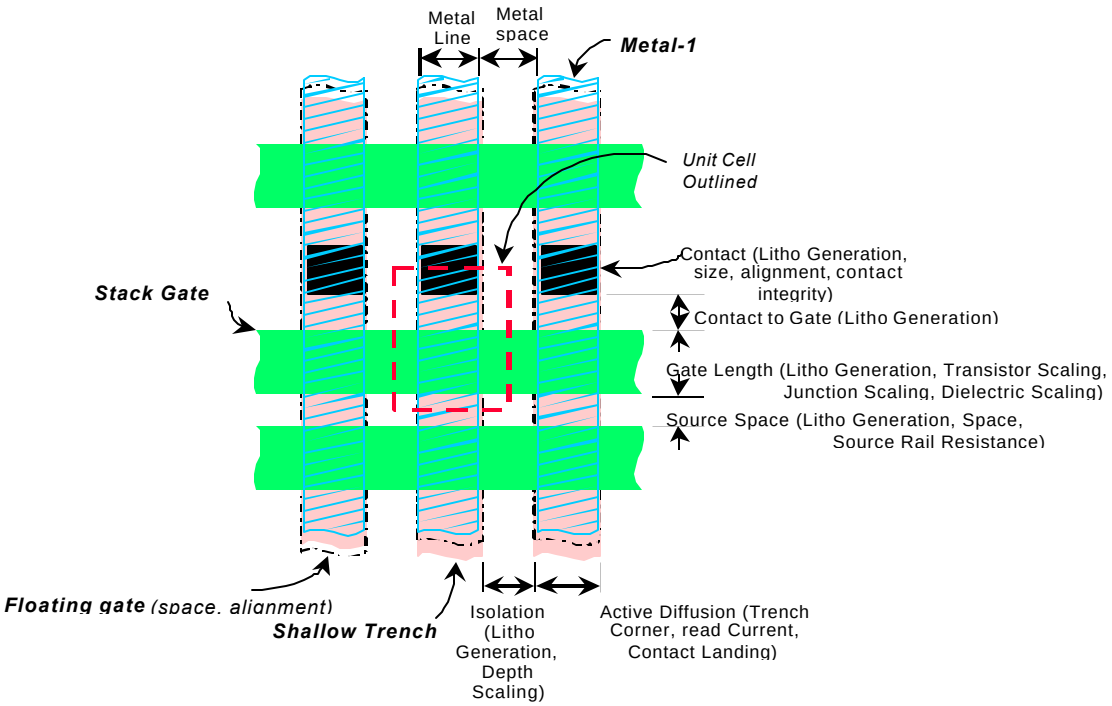


Figure 2: Cell layout and scaling constraints

Contact Size

The key determinants to contact scaling have been the advances in lithography tools, resists, and masks. These have enabled the printing of smaller contacts at every generation. This has been coupled with advances in contact etch chemistry along with the adoption of salicided junctions starting at the 0.25 m generation, eliminating the need for plug implants, required by non-salicided contact processes. The contact plug uses PVD Ti/CVD TiN adhesion layers and blanket tungsten deposition followed by chemical-mechanical polish. The unlanded contact process introduced at 0.18 m (Figure 3, right half) improved registration by allowing a direct contact-to-gate alignment without worrying about alignment to the isolation.

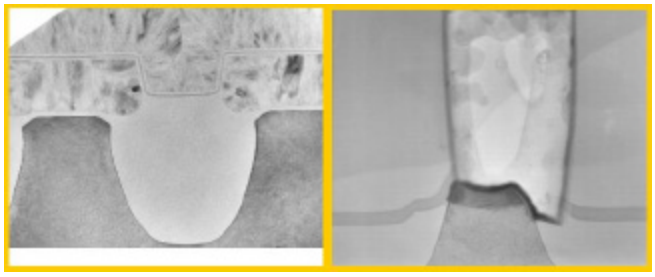


Figure 3: Self-aligned poly and unlanded contact

Source Space Scaling

The primary challenge to scale the source space is to meet the source resistance requirements for each generation. Similar to the contact, the most advanced lithography is used to define the poly space at each generation. A self-aligned source architecture was introduced in the 0.6 m node to eliminate the registration component of the flash cell gate to the diffusion edge, and this continues to be used today. To prevent the source resistance from increasing beyond the maximum requirement, the trench profile and source implants are carefully engineered to manage the trench sidewall resistance without the need for angled implants. The adoption of a dual trench scheme in the 0.13 m generation allowed a much shallower trench to be chosen for the flash array, which made it easier to dope the sidewall, especially at the tighter pitch.

Gate Length Scaling

Gate length has been scaled at each generation using similar techniques to classical transistor scaling, which include junction and channel doping optimization along with gate oxide scaling. In the case of flash, both the tunnel oxide thickness and the interpoly Oxide-Nitride-Oxide (ONO) thickness are scaled to improve the gate coupling to the channel to allow further channel length scaling. In the 0.13 m generation, the ONO effective electrical thickness is 15nm, and the tunnel oxide thickness is 9nm. Changes to the erase scheme have also aided in channel length scaling by allowing the source junction to be scaled, thereby reducing the source junction underlap. At the 0.4 m generation a negative gate erase scheme was adopted, which reduced the cell source voltage from 12V

in the source erase scheme used in earlier generations to ~5V with negative gate erase. At the 0.13 m generation a channel erase scheme was adopted so that the junction could be scaled further as it now no longer needs to support a voltage above the well voltage.

Drain Space Scaling

Generally the drain space is not limited by lithography, as it is larger than the source space, due to the presence of a contact. The key concerns with drain space scaling are adequate contact-to-gate space, which is reduced with improved registration, Inter-Layer-Dielectric gap fill (a HDP oxide is used at the 0.18 m generation and beyond), and the spacer architecture. In the transition from 0.18 m to 0.13 m, a dual spacer scheme was adopted that allowed the flash array spacer to be independent from the periphery high-voltage transistors. This enabled a narrower spacer in the flash drain region so that gap fill was not an issue.

SCALING LIMIT PROJECTION

One can extrapolate the scaling trend based on what has been accomplished so far and the result is shown in Figure 4. This extrapolation is based on the fact that the basic planar cell structure is the same for all the generations, and scaling is achieved by reducing specific cell dimensions. The active electrical cell area is Zeff x Leff, which represents the minimum area required for cell functionality. The trend was relatively flat from 1.0 m to 0.40 m nodes, but was scaled aggressively since the 0.25 m generation. Zphy and Lphy represent the active width and gate length dimensions defined lithographically. The difference between Zphy and Zeff is the beak of the isolation process while the difference between Lphy and Leff is the lateral diffusion of the source and drain underneath the gate. Zphy x Lphy is scaling down at a faster rate compared to Zeff x Leff because of the aggressive reduction of beak and source/drain underlap. However, the beak and source/drain underlap cannot go to zero. Thus, the convergence point of the trend represents a projection of a scaling rate limiter of the current planar cell structure. The trend shows convergence at 45nm, which means that this component of scaling is no longer available. A practical limit of scaling of this component is the 65nm node. This also agrees well with analyses based on other considerations. To continue scaling at the same rate, i.e., meeting Moore's Law, more revolutionary ideas will be needed to either scale the Leff and Zeff more aggressively, which is historically difficult due to hot electron programming limitations, or to go to other cell structures that are not planar (3D cell structures).

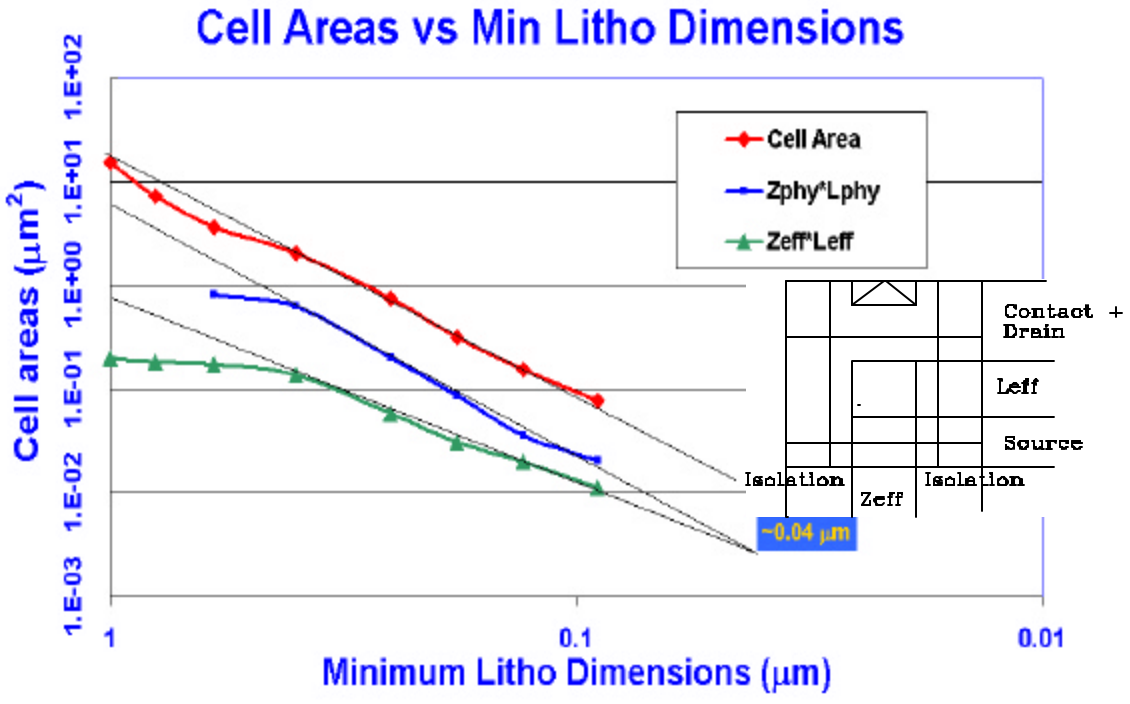


Figure 4: Cell scaling projection

INTEL STRATAFLASH® MEMORY

The Intel StrataFlash memory technology represents a cost breakthrough for flash memory devices by enabling the storage of two bits of data in a single flash memory transistor. Cost-per-bit reduction of flash memory devices has been traditionally achieved by aggressive scaling of the memory cell transistor using silicon process-scaling techniques as discussed in the previous sections of this paper. In an attempt to accelerate the rate of cost reduction beyond that achieved by process scaling, a research program was started in 1992 to develop methods for the reliable storage of multiple bits of data in a single flash memory cell. The result of this research was the commercial introduction of the first Intel StrataFlash memory in 1997, utilizing the 0.4 m technology node. The two-bit-per-cell Intel StrataFlash memory technology provides a cost structure equivalent to the next generation of process technology while using the current generation of process technology equipment. Today, the Intel StrataFlash memory technology has become the mainstream flash solution.

The Multi-Bit Storage Breakthrough: Intel StrataFlash® Memory Technology

As discussed earlier, the flash memory device is a single transistor that includes an isolated floating gate. The floating gate is capable of storing electrons. The behavior of the transistor is altered depending on the amount of charge stored on the floating gate. Charge is placed on the floating gate through a technique called programming. The programming operation generates hot electrons in the channel region of the memory cell transistor. A fraction of these hot electrons gain enough energy to surmount the 3.2eV barrier of the Si-SiO_2 interface and become trapped on the floating gate. For single-bit-per-cell devices, the transistor either has little charge (<5,000 electrons) on the floating gate and thus stores a "1," or it has a lot of charge (>30,000 electrons) on the floating gate and thus stores a "0." When the memory cell is read, the presence or absence of charge is determined by sensing the change in the behavior of the memory transistor due to the stored charge. The stored charge is manifested as a change in the threshold voltage of the memory cell transistor. Figure 5 illustrates the threshold voltage distributions for a half-million cell (1/2Mc) array block. After erasure or programming, the threshold voltage of every memory cell transistor in the 1/2Mc block is measured, and a histogram of the results is presented. Erased cells (data 1) have threshold voltages less than 3.1v, while programmed cells (data 0) have threshold voltages greater than 5v.

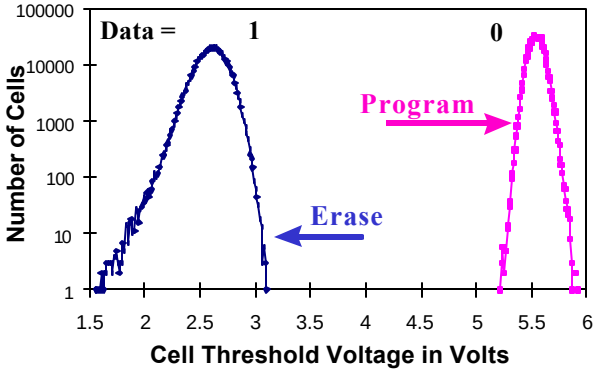


Figure 5: Single-bit/cell array threshold voltage histogram

The charge storage ability of the flash memory cell is a key to the storage of multiple bits in a single cell. The flash cell is an analog storage device, not a digital storage device. It stores charge (quantized at a single electron), not bits. By using a controlled programming technique, it is possible to place a precise amount of charge on the floating gate. If charge can be accurately placed to one of four charge states (or ranges), then the cell can be said to store two bits. Each of the four charge states is associated with a two-bit data pattern. Figure 6 illustrates the threshold voltage distributions for a 1/2Mc block for two bits per cell storage. After erasure or precise programming to one of three program states, the threshold of each of the 1/2Mc is measured and plotted as a histogram. Notice the precise control of the center two states, each of which is approximately 0.3v (or 3,000 electrons) in width.

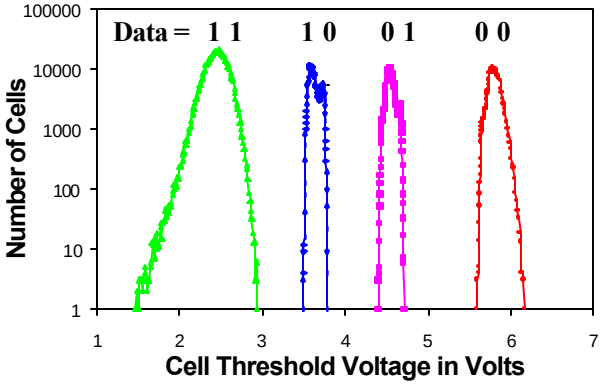


Figure 6: Two-bit/cell array threshold voltage histogram

Higher bit-per-cell densities are possible by even more precise charge placement control. Three bits per cell require eight distinct charge states and four bits per cell

require sixteen distinct charge states. In general, the number of states required is equal to 2^N where N is the desired number of bits.

The ability to precisely place charge on the floating gate and at some later time sense the amount of charge that was stored has required substantial innovations, and extensive characterization and understanding of cell device physics, memory design, and memory test. These innovations are discussed in detail in two earlier *Intel Technology Journal* papers [4,5].

LOW-VOLTAGE, HIGH-PERFORMANCE OPTIMIZATION FOR DISCRETE FLASH MEMORIES

Increasing read performance demands at low operating voltage taxes the ability of high-voltage transistors, which are required by flash program and erase. Cobalt-salicided complementary polysilicon gates are used to form low-threshold NMOS and PMOS surface-channel transistors and low source/drain and gate resistance. Additionally, special low-threshold devices, for low-voltage performance and analog circuit design requirements, are provided by separate well and Vt-adjust implants. Continual application performance demands and further reductions in operating voltages require the inclusion at the 0.18 m technology generation of thin gate-oxide logic compatible NMOS and PMOS transistors. This is achieved with three additional masking layers (one for thin gate oxide and two for low-voltage wells). Source/drain and tip regions are shared between the low-voltage and high-voltage transistors to best balance performance with added processing steps. The thin gate-oxide architecture is bounded by optimization for low voltage (<1.8V), while maintaining compatibility for legacy voltage (3.3V), including balancing of device Vt with off-current leakage for minimization of standby currents. An 8nm gate oxide was chosen to balance these needs, with trench processing meeting charge-to-breakdown requirements, supporting three separate oxides: tunnel oxide, high-performance oxide, and high-voltage oxide. A triple well is provided for design flexibility of negative voltage switching and low-voltage optimization. Lastly, performance capability is provided by three layers of aluminum metalization, allowing additional wordline and bitline strapping of the flash array, for reduced resistance-capacitance, RC delay, and more efficient signal routing in the periphery.

In addition to low voltage and high performance, the trench isolation, thin-gate-oxide, salicided complementary poly gate transistors and the three layers of interconnect inclusion provide all the key architectural elements required for embedded logic capability. Higher degrees of thin-gate device performance can be achieved by further separating the process steps and reducing the oxide thickness for lower voltage operation, as discussed below.

Lastly, the cost sensitivity of the market for memory dictates requirements for low-cost process technologies. The described cell scaling and Intel StrataFlash memory capability satisfy low cost. Additionally, process synergy of this memory process technology, with the basic process modules and equipment set with other high-volume logic technologies, lower cost through economies of scale by providing factory flexibility and shared process step and yield learning.

To reduce cost, the periphery transistors must also be scaled since they constitute a significant portion of the die area. The introduction of channel erase reduces the maximum voltage the periphery needs to support, and the introduction of more advanced lithography and etch gives better gate-patterning capability. These allow the channel length and gate oxides to be scaled, which is done in conjunction with traditional junction scaling, and which leads to a significant reduction in the gate length, while at the same time maintaining good transistor characteristics. For the embedded logic process, below, this leads to a gate length of 100nm. The reduction in the maximum voltage the periphery needs to support along with the dual trench scheme allows the isolation width to be scaled as well, since a deep trench can be maintained for logic devices independent of the shallow trench used in the flash array. These changes, combined with the advanced 0.13 m lithography tools, cobalt salicide, and complimentary gates consistent with Intel's 0.13 m logic process, deliver the required transistor performance and area savings.

WIRELESS INTERNET ON A CHIP

Traditionally, flash and logic process technologies are optimized separately on separate process equipment sets and separate fabrication facilities. During the development of the 0.25 m flash process technology, Intel made the strategic decision to develop its flash processes synergistically with its logic processes. This initial decision was made with the goal of processing the two technologies, flash and logic, in the same fabrication facility, for improved manufacturing flexibility and shared learning and for maximum volume production efficiency. This decision also brought key process modules into the flash processes, which historically were not found on flash, such as trench isolation and salicided

complimentary gates. Both of these process modules are examples of key enablers that not only achieve the manufacturing synergy goal, but also provide for dramatic advancement of scaling the flash memory cell and enhancing performance. (This was outlined in the previous sections.) Additionally, the incorporation of these features into flash memory technology has paved the way for the integration of a high-performance logic function with a dense flash memory on the same chip. This capability has led to the "wireless Internet on a chip" technology, where all the key elements of a typical cell phone and a typical handheld computer, the advanced digital logic functions, all the SRAM and flash memory functions, and the analog functions for interfacing to a radio are all integrated onto a single chip. This is cost-effectively achieved without compromising the performance of the state-of-the-art digital logic or the density of the state-of-the-art flash memory.

The value of this integration is several fold. First, the total number of devices can be reduced, thereby reducing the form factor of a wireless device, allowing for smaller lighter devices. What were previously several chips is now reduced to one. The reduction in the number of chips in a system also improves overall system reliability. Secondly, the integration of flash memory serves to enhance the performance of the digital logic computing functions. Memory latency is greatly reduced, and bandwidth is greatly enhanced by having logic and memory functions integrated onto the same chip. Lastly, this enhanced performance is achieved at lower power, as interconnect bus capacitance is significantly reduced with an integrated on-chip bus, versus a discrete external bus.

Five key innovations are required to achieve the "wireless Internet on a chip" technology. They are the key process modules for advanced logic functionality with an advanced flash process. These five innovations are trench isolation, a multiple gate-oxide process, a low thermal budget, salicided complementary gates, and a multi-level metal system.

Trench isolation is required for tight pitch logic design rules, for high transistor count design, and for small SRAM memory cell layout. Trench isolation is not typically found on flash processes, due to the challenges outlined earlier. These challenges were overcome with the integration of trench isolation in the 0.25 m flash process and have served as the basis for cell size reduction in the flash cell.

Multiple gate oxides are required to achieve the separate function required for the high-voltage operation of the flash cell and the ultra-low voltage required for the logic operation. The 0.18 m flash process incorporated multiple periphery gate oxides, as outlined earlier. This same process architecture is extended to achieve the ultra-thin (<3nm) gate oxide required for advanced logic functions.

A low thermal budget processing is required for high-performance transistors. Traditionally, memory plus logic integrated with memories such as DRAMs have had difficulties with achieving a low thermal budget, as the DRAM cell processing (requiring high temperatures) is often done subsequent to the formation of the logic transistors, thereby significantly limiting the performance of the logic functions. This is not the case with flash memory integration, as the flash memory processing occurs earlier than the formation of the logic transistors. As such, the high-thermal process steps of the flash memory are not seen by the logic transistors, thereby maintaining the high-performance capability of the logic functions.

Salicided complementary gates are required to achieve low-threshold voltages and short channel lengths that are required for high-performance logic functions. Salicided gates are often difficult to integrate with memories, as tight spaces found in memories pose challenges to salicide processing. These barriers were overcome in the 0.25 m node flash technology, with the integration of salicide, outlined earlier.

Lastly, multiple metal layers are required for high transistor-count logic designs. The metal processing is accomplished with backend planarization, fully compatible with the logic and flash processing.

With these innovations, the ability to fully integrate state-of-the-art logic performance and state-of-the-art flash memory density, cost effectively, without compromising either, has been fully realized. The analog features are relatively simple process components, most of which are found in standard flash processing. Key attributes for analog processing are a triple well for noise isolation that is standard in flash memories, 3V optimized transistors that are also standard in flash memories, and precision resistor and capacitor passives, that can be bolted on, relatively simply, to a CMOS process.

CONCLUSION

By following Moore's law, ETOX™ flash memory has gone from 1.5 m node in development in the mid 1980s to 0.13 m in high-volume production today. The scaling has been accomplished by improved lithography capability as

well as many innovations. In this paper, we reviewed key scaling challenges as well as the key innovations. Based on projection, the current planar cell structure can be scaled to the 65nm node. More revolutionary innovation such as 3D structures may be required for the 45nm node and beyond. To lower cost further, we have developed the Intel StrataFlash® memory technology, which stores two bits of information in a single physical memory cell. The scaling capability also allows for the integration of flash memories with high-performance logic for "wireless Internet on a chip" technology.

ACKNOWLEDGMENTS

The authors acknowledge the members of Intel's California Technology and Manufacturing organization and the Flash Products Group for their efforts over the past eight generations of technologies.

REFERENCES

[1] S. Lai, "Flash Memories: Where We Were and Where We Are Going," *IEEE IEDM Tech. Digest*, 1998, pp. 971-3.

[2] A. Fazio, "A High Density High Performance 180nm Generation High Density Etox™ Flash Memory Technology," *IEEE IEDM Tech. Digest*, 1999, pp. 267-270.

[3] S. Keeney, "A 130nm Generation High-Density Etox™ Flash Memory Technology," *IEEE IEDM Tech. Digest*, 2001, pp. 41-44.

[4] G. Atwood, et. al., "Intel StrataFlash Memory Technology Overview" *Intel Technology Journal*, Q4, 1997 at http://developer.intel.com/technology/itj/q41997/articles/art_1.htm

[5] A. Fazio, et. al., "Intel StrataFlash Memory Development and Implementation" *Intel Technology Journal*, Q4, 1997 at http://developer.intel.com/technology/itj/q41997/articles/art_2.htm

AUTHORS' BIOGRAPHIES

Al Fazio is a Principal Engineer responsible for Communication Technology Development. He joined Intel in 1982 after receiving his B.S. degree in Physics from the State University of New York at Stony Brook. He has worked on numerous memory technologies and was responsible for the Intel StrataFlash memory and Flash+Logic+Analog "wireless Internet on a chip" technology developments. Al holds over 20 patents and has written several technical papers, two of which have won outstanding paper awards at IEEE-sponsored conferences. He has served as general chairman of the IEEE NVSMW. His e-mail is al.fazio@intel.com

Stephen Keeney is the Process Integration Manager for the 0.13 m and 0.09 m flash technology development programs. Stephen obtained his B.E. degree from University College Dublin, Ireland in 1988 and his Ph.D. degree in Microelectronics from the NMRC, Cork, Ireland in 1992. He joined Intel in 1993 and has worked extensively across many aspects of flash memory development, including device physics innovations; yield analysis, memory test architecture, process integration and Intel StrataFlash memory. Stephen holds six patents and has written 20 technical papers. His e-mail is stephen.n.keeney@intel.com

Stefan K. Lai is Vice President, Technology and Manufacturing Group, and Director, California Technology and Manufacturing. Stefan is responsible for the development of silicon process technologies for devices used in communications products, including flash, flash+logic, analog, and novel memory technologies. He was recognized as an IEEE Fellow in 1998 for his research on the properties of silicon MOS interfaces and the development of flash EPROM memory. Stefan received a B.S. degree in Applied Physics from the California Institute of Technology in 1973, and a Ph.D. degree in Applied Quantum Physics from Yale University in 1979. He joined Intel in 1982. His e-mail is stefan.lai@intel.com

For further information visit:

developer.intel.com/technology/itj/index.htm